Electronic Supplementary Material (ESI) for Journal of Analytical Atomic Spectrometry.

Soft classification of single samples based on multi-analyte spectra – Electronic supplementary information

Nai-Ho Cheung

Department of Physics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China

Abstract

The published article reports the *soft* classification of Chinese red seal inks based on the laser-induced fluorescence spectra of their plumes. Some details of the computation that are supplementary to the main arguments are presented here. They include (1) the projection from *y* to *z* space, (2) the derivation of similarity in naïve Bayes probabilistic language, (3) the determination of the no-class threshold $\underline{S}$, (4) the optimization of the similarity range parameter β, (5) the differentiating and non-differentiating spectral features, (6) the determination of the no-class probability, and (7) the evaluation of the statistical figures of merit.

List of Figures

List of Tables

From *y* space to *z* space

In the Methods section of the article, we used the sorting of three inks, C2, C4 and C5 to illustrate our soft classification scheme. We showed in Fig. 1c of the article that all the data points lie on a two-dimensional plane defined by the three anchor points (y_2, y_4, y_5): (1,0,0), (0,1,0) and (0,0,1). In Fig. 2a, we labeled the coordinates of that plane $z_1$and z_2. The linear dependence among the *y* variables, $y_2 + y_4 + y_5 = 1$, can be side-stepped by using (z_1, z_2) instead of (y_2, y_4, y_5) to represent the data points.

The transformation from ***y*** to ***z*** is best carried out by doing partial-least-square discriminant-analysis (PLSDA) with the ***y*** data as input. The latent variables will be our desired ***z***. For the case of 3-class sorting, we will have only two latent variables, as expected.

An equivalent method that offers a clearer geometrical picture is to treat the ***y*** to ***z*** projection as a coordinate rotation. It is illustrated in Fig. E-1. The data plane is represented by the blue equilateral triangle with vertices at (1,0,0), (0,1,0) and (0,0,1). A second equilateral triangle, gray in color, is constructed to be identical to the blue but offset rearwards to have its centroid coincides with the origin of the *y* coordinate system. The z_1 axis is drawn through the centroid and parallel to the base of the gray triangle. The z_2 axis is drawn through the centroid and perpendicular to z_1. The z_3 axis (not shown in Fig. E-1) points from the origin to the centroid of the blue triangle.

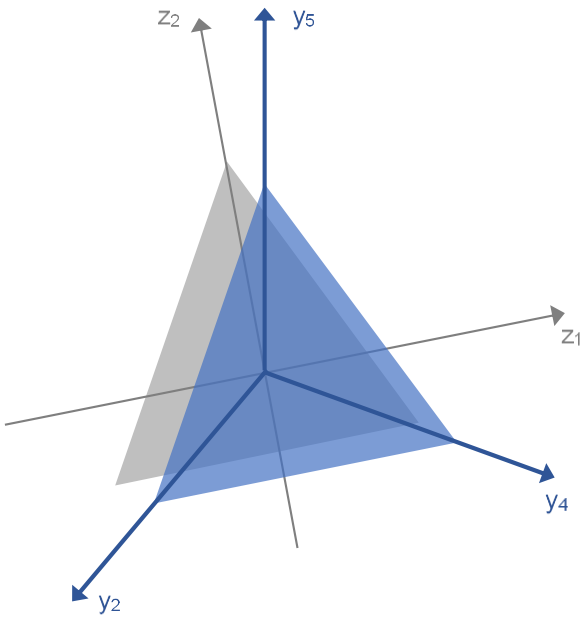

Fig. E-1. The blue equilateral triangle represents the data plane in the y_2-y_4-y_5 class space. The gray equilateral triangle is identical to the blue but offset rearwards to have its centroid coincides with the origin of the *y* coordinate system. The z_1 axis is parallel to the base of the gray triangle and passes through the centroid. The z_2 axis is orthogonal to z_1 and also passes through the centroid.

As can be seen from Fig. E-1, the ***y*** to ***z*** transformation is a rotational transform, ***z*** = **R*y***. The matrix elements of **R** can be determined in two steps.

Step 1. We can see that z_3 of all data points will be $1/\sqrt{3}$, the distance between the origin and the centroid of the blue triangle. By requiring $z_3 = 1/\sqrt{3}$ for the three anchor vectors (the three vertices of the blue equilateral triangle),

$$y_2 = \begin{bmatrix}1\\0\\0\end{bmatrix}, \qquad y_4 = \begin{bmatrix}0\\1\\0\end{bmatrix}, \qquad y_5 = \begin{bmatrix}0\\0\\1\end{bmatrix}, \tag{E-1}$$

we can see that $R_{31} = R_{32} = R_{33} = 1/\sqrt{3}$.

Step 2. Each row of **R** can be treated as a row vector. Because **R** is a rotational transform, it has to be orthogonal. The three row vectors therefore form a set of orthonormal vectors. The bottom row vector is known from Step 1, namely, $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$. The other two row vectors can be found by Gram-Schmidt orthogonalization.

Generalizing to *d* class sorting, each entry of the bottom row of **R** can be shown to be $1/\sqrt{d}$. The **R** matrices for three and five -class sorting are shown in Fig. E-2.

$$\begin{bmatrix} 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \sqrt{\frac{2}{3}} & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \qquad \begin{bmatrix} 0 & 0 & 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 & \sqrt{\frac{2}{3}} & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} \\ 0 & \frac{\sqrt{3}}{2} & \frac{-1}{\sqrt{12}} & \frac{-1}{\sqrt{12}} & \frac{-1}{\sqrt{12}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{20}} & \frac{-1}{\sqrt{20}} & \frac{-1}{\sqrt{20}} & \frac{-1}{\sqrt{20}} \\ \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix}$$

Fig. E-2. The ***y*** to ***z*** rotational transforms for three (left) and five (right) dimensions.

It should be noted that the matrix elements (Fig. E-2) are not unique. For example, in the case of 3-class sorting, any rotation about z_3 will produce a valid z_1-z_2 plane. The particular ***z*** plane generated by PLSDA corresponds to maximum variance along z_1, then the next maximum variance along z_2, etc. Similarity functions constructed in this kind of ***z*** space will best approximate the cluster distributions. We therefore use the PLSDA method to transform ***y*** to ***z***.

A final word of caution. As mentioned in the article, three-class sorting should be based on similarity *S* that is measured in the two-dimensional z_1-z_2 space. Similarity that is

measured in two-dimensional y_2-y_4 space (or any two of the three y's) is wrong. The reason can be made obvious by considering the following example. Suppose the C2, C4 and C5 training clusters center on (y_2, y_4, y_5) = (1,0,0), (0,1,0) and (0,0,1), respectively. Now suppose we have two test points, #1 at (-0.5, 0.5, 1) and #2 at (-0.5, 1.5, 0). Their respective distance from the C4 cluster is $\sqrt{1.5}$ and $\sqrt{0.5}$, i.e., #1 being less similar. However, in y_2-y_4 space, they are equidistant from C4 and their similarities so computed will be identical.

Naïve Bayes and Similarity

In section 2.2 of the article, we gave an intuitive explanation of the naïve Bayes (NB) classifier in terms of similarity. Here, we will present NB in probability formalism and illustrate how probabilities are related to similarities.

We are to classify an unknown ink sample in $\boldsymbol{z}$ space. Before we measure the PLIF spectrum of the unknown, the *joint* probability $P(C, \boldsymbol{z})$ that (1) the unknown belongs to ink class C and (2) its PLIF spectrum maps to class space coordinate $\boldsymbol{z}$ is given by,

$$P(C, \boldsymbol{z}) = P(C|\boldsymbol{z})P(\boldsymbol{z}), \quad \text{(E-2)}$$

$$= P(\boldsymbol{z}|C)P(C), \quad \text{(E-3)}$$

where $P(C|\boldsymbol{z})$ and $P(\boldsymbol{z}|C)$ are *conditional* probabilities, and $P(\boldsymbol{z})$ and $P(C)$ are *prior* probabilities.

Once we measured its PLIF spectrum and determined its coordinate to be $\boldsymbol{Z}$, the prior $P(\boldsymbol{Z}) = 1$, and Eqs. (E-2) and (E-3) give,

$$P(C|\boldsymbol{Z}) = P(\boldsymbol{Z}|C)P(C). \quad \text{(E-4)}$$

We will assume that the prior $P(C)$ is a constant independent of class, i.e., we assume, *a priori*, that ink samples have equal chance to belong to any ink class. Eq. (E-4) then becomes,

$$P(C|\boldsymbol{Z}) \propto P(\boldsymbol{Z}|C). \quad \text{(E-5)}$$

The LHS of Eq. (E-5) is our *prediction* task. It says, once we PLIF-analyzed the unknown and mapped the spectrum to class space coordinate $\boldsymbol{Z}$, what is the class C membership probability? The answer is given by the RHS of Eq. (E-5).

What is the RHS of Eq. (E-5) saying? It is the chance that a class C sample has coordinate $\boldsymbol{Z}$. Do we know it? Well, suppose we have one *training* sample of class C with coordinate $\boldsymbol{z}$. We can then approximate the RHS of Eq. (E-5) by the so-called Gaussian kernel,

$$P(\boldsymbol{Z}|C) = exp\left[-\frac{1}{2}\left(\frac{\boldsymbol{Z}-\boldsymbol{z}}{\delta}\right)^2\right], \quad \text{(E-6)}$$

where δ is a measure of the uncertainty in $\boldsymbol{z}$. Eq. (E-6) is identical to Eq. (1) of the article, and $P(\boldsymbol{Z}|C)$ is a measure of the similarity of the unknown and the training sample.

Determination of the no-class threshold $\underline{S}$

The no-class threshold $\underline{S}$ is given by the similarity of a boundary point that is closest to the cluster. For 2-dimensional $\boldsymbol{z}$ space and 99.7% inclusion, the boundary points lie on the ellipse defined by $\mathbf{z}_c \pm 3.18\boldsymbol{\sigma}$ (see Fig. 2a of the article). To find $\underline{S}$, we generate data points randomly on the ellipse and compute their S. The maximum S among them will approximate $\underline{S}$ very well if we generate enough data points. We found that 30 or more data points will be adequate for sampling the 2-dimensional ellipse. For 4-dimensional $\boldsymbol{z}$ space, at least 30^3 = 27,000 sampling points are needed. The values of the thresholds, one for each of the five ink classes, are listed in Table 1 of the article.

The optimization of the similarity range β

As explained in the article, for a test sample at position **z**, its similarity to a nearby training cluster is given by the scalar field $S(\boldsymbol{z})$ in a d-dimensional **z** space. $S(\boldsymbol{z})$ can be visualized as the summation of little ellipsoids, one on each of the n training observations, with radius δ_j along y_j given by,

$$\delta_j = \frac{\beta \sigma_j}{\sqrt[d]{n}}, \qquad \text{(E-7)}$$

where σ_j is the cluster spread (standard deviation) along z_j and β is a constant whose value is to be optimized. A larger β means bigger ellipsoids to smooth out $S(\boldsymbol{z})$, but too large a β will invite interference by neighboring clusters.

We can use the C5 cluster in 4-dimensional **z** space as an example. We set β equals to 1 and plot $S_5(\boldsymbol{z})$ for various **z** positions: from the centroid of the cluster to an outermost training observation. We then repeat with β = 2, 3.. etc. The results are shown in Fig. E-3. As can be seen, for β too small, $S_5(\boldsymbol{z})$ is grainy. Only when β > 4 will the statistical fluctuations be smoothed out.

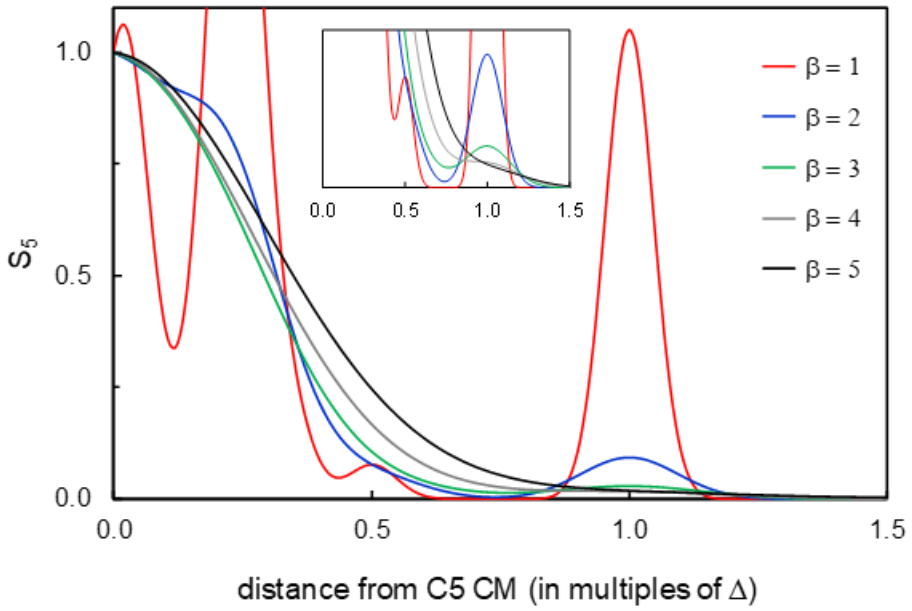

Fig. E-3. C5 similarity function at various locations from the centroid (CM) of the C5 training cluster to an outermost training observation and beyond, for five values of β. The separation of the CM and the outermost observation is defined as one unit. Inset shows S_5 magnified 10×.

With increasing β, the similarity envelope extends further out so S_i for test observations of class *i* will increase. S_i will approach unity asymptotically. This is shown in Fig. E-4 (red data points) when $\langle S_i \rangle$ of the 92×5 test observations is plotted against β. At the same time, interference by neighboring clusters, $\langle \sum_{k \neq i} S_k \rangle$, will also increase. This is shown by the blue data points in Fig. E-4. As can be seen, interference increases sharply when β is greater than 5.

We therefore set β = 5 in all subsequent calculations. Accordingly, the ellipsoid radius is about 1.36× the cluster spread for the case of 184 training observations in 4-dimensional ***z*** space (Eq. 5 of the article, with d = 5 and n = 184).

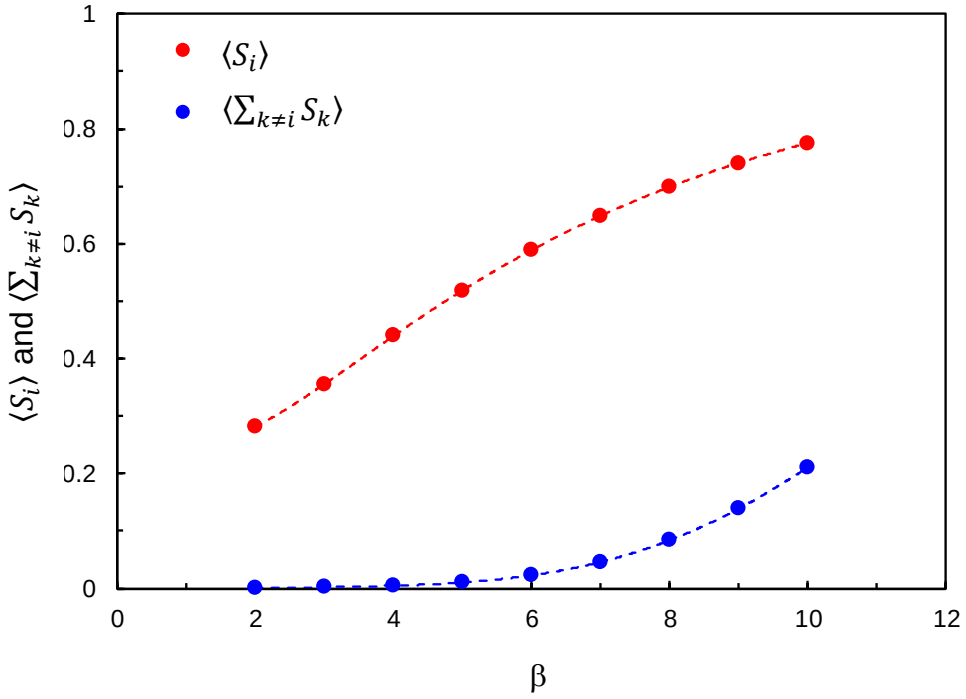

Fig. E-4. Plot of $\langle S_i \rangle$ and $\langle \sum_{k \neq i} S_k \rangle$ against β. Dashed lines are visual aids.

Differentiating and non-differentiating features

We showed in the article that the differentiating features are given by the predictive component of the variable-importance-in-projection (VIP) while the non-differentiating features are given by the orthogonal component of the VIP.

We can also deduce the differentiating features by a more intuitive approach. We first compute the class-averaged spectra $I_i(\lambda)$ for class C_i. We then average all five $I_i(\lambda)$ to produce the global averaged spectra $I(\lambda)$. The difference spectra, $\Delta_i(\lambda) = |I - I_i|$, indicates how ink *i* differs from the norm. The class-averaged $\Delta_i(\lambda)$, $\langle\Delta(\lambda)\rangle$, is therefore a spectrum of the differentiating features. It is shown in Fig. E-5 (blue upper trace). As can be seen, it resembles the predictive VIP (red lower trace). Their consistency lends support to our interpretation of the predictive VIP.

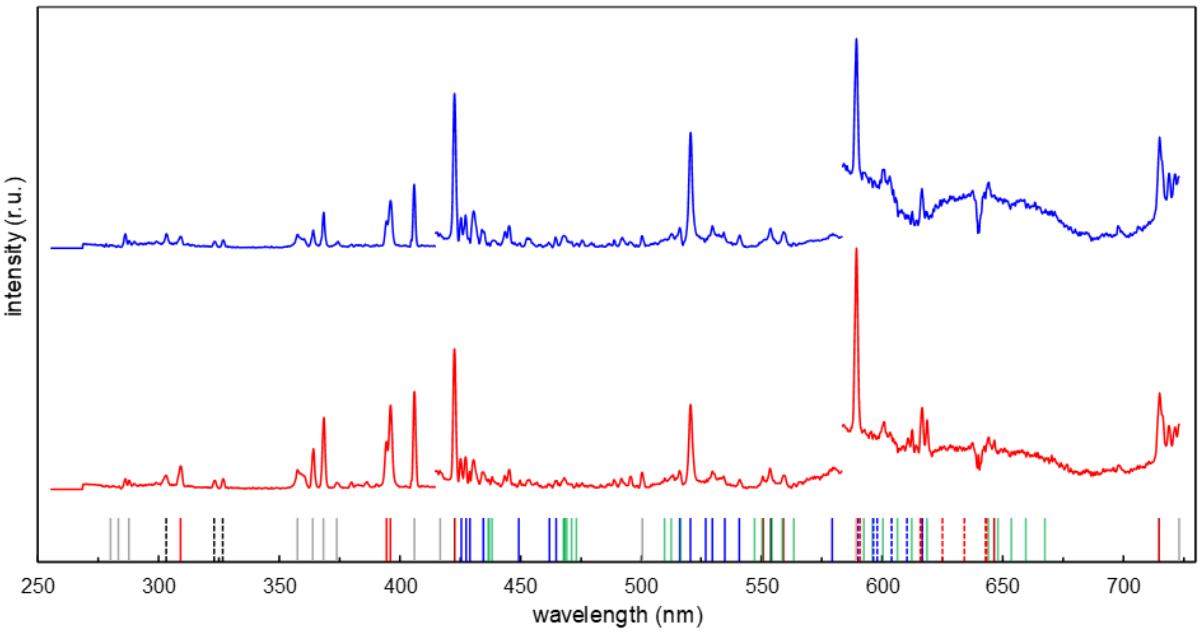

Fig. E-5. The spectrum of the predictive VIP, shown in red; and the $\langle\Delta(\lambda)\rangle$ spectrum, shown in blue. The two spectra are offset vertically for clarity, their leading pixels are zeroed to indicate the baselines, and their spectral intensities are normalized to the same scale. Identities of the stronger lines are color-coded at the bottom: Al I (red), Ba I (black), Ca I (brown), Cr I (blue), Na I (orange), Pb I (gray), Sb I (black, dotted), C_2 band heads (green), S_2 band heads (blue, dotted) and PbO band heads (red, dotted).

Similarly, we can guess the non-differentiating features by simply computing the standard deviation spectra $\sigma_i(\lambda)$ for class C_i and averaging all five $\sigma_i(\lambda)$ to produce $\langle\sigma(\lambda)\rangle$. This mean-deviation spectrum $\langle\sigma(\lambda)\rangle$ represents the intra-class variations. It is shown in Fig. E-6 (blue upper trace). Again, its resemblance to the orthogonal VIP spectrum (red lower trace) is evident. It validates our interpretation of the orthogonal VIP.

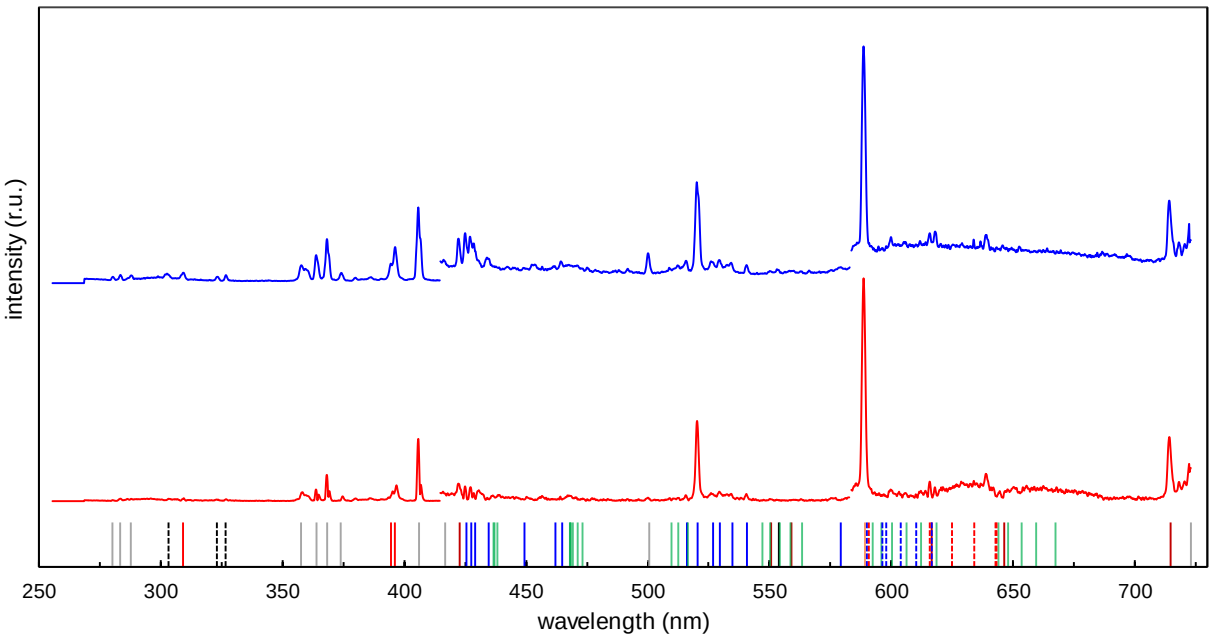

Fig. E-6. The spectrum of the orthogonal VIP, shown in red; and the average σ spectrum, shown in blue. The two spectra are offset vertically for clarity, their leading pixels are zeroed to indicate the baselines, and their spectral intensities are normalized to the same scale. Identities of the stronger lines are color-coded at the bottom, as explained in the caption of Fig. E-5.

Computing no-class probability

In the article, we classified each and every C6 sample as no-class without giving the numerical no-class probability. It is possible to evaluate the no-class probability P_{NC} by going through the following steps.

1. Use the 99.7% inclusion threshold, identify the NC samples.
2. Then set the % inclusion to 99.8% (always ≥ 99.7%) and compute the corresponding $\underline{S}_i$. Then set % inclusion to 99.9% and compute the corresponding $\underline{S}_i$ again, etc. Then plot % inclusion (along y) against $\underline{S}_i$ (along x) and functionally fit the curve to give $y(x)$, as shown in Fig. E-7 for % inclusion versus $\underline{S}_1$. Note that % inclusion has the meaning of P_{NC}. For example, 99.8% inclusion implies that the probability of being a non-member is 99.8%.
3. Now, for that NC sample, based on its S_i values, we can compute the corresponding P_{NC} for each class C_i.

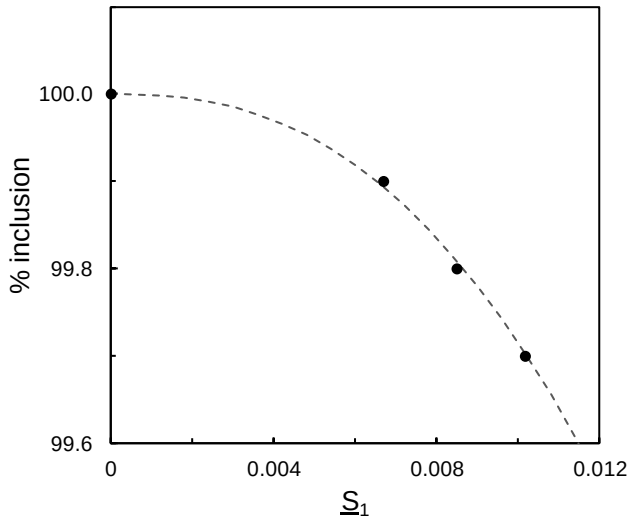

Fig. E-7. % inclusion against $\underline{S}_1$.

We follow the procedure and evaluate the P_{NC} of the C1 sample that was falsely classified as no-class (see Table 2 of the article). The results are listed in Table E-1. As can be seen, obs # 958 is remotely similar to its own class.

Table E-1. P_{NC} of the C1 sample that was wrongly classified as no-class.

		P_{NC} (%)				
Obs #	Class	C1	C2	C3	C4	C5
958	C1	99.89	100	100	100	100

Statistical figures of merit

The figures of merit listed in Table 4 of the article can be readily computed from Tables 2 and 3 of the article by following these steps.

1. For each of the two schemes, create a Positive Table. For the hard scheme, it is the same as its confusion matrix but with the C6 row left out. For the soft scheme, an equivalent confusion matrix can be drawn based on Table 2 of the article. Results are shown in Tables E-2 and E-3.
2. Create a Negative Table for each scheme, with each entry = (92 – corresponding entry of its Positive Table). Results are shown in Tables E-4 and E-5.
3. For each positive table, the maximum positive counts P = sum of all elements of the Positive Table = 460 for both hard and soft schemes.
4. For each negative table, the maximum negative counts N = sum of all elements of the Negative Table = 2,300 for both schemes.
5. The True Positive counts TP = sum of all diagonal elements of the positive table. TP = 444 and 455 for hard and soft, respectively.
6. The False Positive counts FP = sum of all off-diagonal elements of the positive table. FP = 16 and 5 for hard and soft, respectively.
7. The True Negative counts TN = sum of all off-diagonal elements of the negative table. TN = 2,284 and 2,295 for hard and soft, respectively.
8. The False Negative counts FN = sum of all diagonal elements of the negative table. FN = 16 and 5 for hard and soft, respectively. Can see that $FP = FN$.
9. The False No-Class counts FNC = sum of all elements under the NC column of the positive table. FNC = 0 and 1 for hard and soft, respectively.
10. The maximum false no-class counts = P = 460.
11. The False In-Class count FIC = # C6 samples sorted as in-class. FIC = 92 and 0 for hard and soft, respectively.
12. The maximum false in-class counts = total number of C6 test samples = 92.

Table E-2. Positive Table of hard scheme.

	C1	C2	C3	C4	C5	NC
C1	91	1	0	0	0	0
C2	0	91	0	0	1	0
C3	0	0	85	7	0	0
C4	3	1	0	88	0	0
C5	0	1	0	2	89	0

Table E-3. Positive Table of soft scheme.

	C1	C2	C3	C4	C5	NC
C1	91	0	0	0	0	1
C2	0	91	0	0	1	0
C3	0	0	92	0	0	0
C4	1	0	0	90	1	0
C5	0	1	0	0	91	0

Table E-4. Negative Table of hard scheme

	Not C1	Not C2	Not C3	Not C4	Not C5	Not NC
C1	1	91	92	92	92	92
C2	92	1	92	92	91	92
C3	92	92	7	85	92	92
C4	89	91	92	4	92	92
C5	92	91	92	90	3	92

Table E-5. Negative Table of soft scheme

	Not C1	Not C2	Not C3	Not C4	Not C5	Not NC
C1	1	92	92	92	92	91
C2	92	1	92	92	91	92
C3	92	92	0	92	92	92
C4	91	92	92	2	91	92
C5	92	91	92	92	1	92